Immunology
The Journal of cells, molecules, systems and technologies

British Society for immunology

IMMUNOLOGY ORIGINAL ARTICLE

Selection influences naive CD8+ TCR-β repertoire sharing

Hao H. Yiu,[1] Louis N. Schoettle,[2] Marlene Garcia-Neuer,[2] Joseph N. Blattman[2] and Philip L. F. Johnson[1]

[1]*Department of Biology, University of Maryland, College Park, MD, USA and* [2]*School of Life Sciences, The Biodesign Institute, Arizona State University, Tempe, AZ, USA*

doi:10.1111/imm.13299
Received 16 January 2020; revised 22 November 2020; accepted 29 November 2020.

Correspondence
Philip L. F. Johnson, Department of Biology, University of Maryland, College Park, College Park, MD 20742, USA.
Email: plfj@umd.edu
Senior author: Philip L. F. Johnson

Summary

Within each individual, the adaptive immune system generates a repertoire of cells expressing receptors capable of recognizing diverse potential pathogens. The theoretical diversity of the T-cell receptor (TCR) repertoire exceeds the actual size of the T-cell population in an individual by several orders of magnitude – making the observation of identical TCRs in different individuals extremely improbable if all receptors were equally likely. Despite this disparity between the theoretical and the realized diversity of the repertoire, these 'public' receptor sequences have been identified in autoimmune, cancer and pathogen interaction contexts. Biased generation processes explain the presence of public TCRs in the naive repertoire, but do not adequately explain the different abundances of these public TCRs. We investigate and characterize the distribution of genomic TCR-**β** sequences of naive CD8+ T cells from three genetically identical mice, comparing non-productive (non-functional sequences) and productive sequences. We find public TCR-**β** sequences at higher abundances compared with unshared sequences in the productive, but not in the non-productive, repertoire. We show that neutral processes such as recombination biases, codon degeneracy and generation probability do not fully account for these differences, and conclude that thymic or peripheral selection plays an important role in increasing the abundances of public TCR-**β** sequences.

Keywords: biased recombination; public; repertoire; selection; sharing; TCR.

INTRODUCTION

The adaptive immune system relies on the generation of a diverse receptor repertoire to maximize the chance of detecting potential pathogens. T-cell receptors (TCRs) mediate detection of pathogens by recognizing peptides presented by the major histocompatibility complex (MHC) on most nucleated cells. The richness of the TCR repertoire (>10^{15} distinct TCRs possible in mice[1]) relative to the number of T cells (estimated 10^7 naive T cells in an individual)[2,3] makes TCRs common to multiple individuals implausible at face value—yet numerous clinically relevant public TCR sequences have been reported.[4] Public TCRs may effectively defend against frequently encountered pathogens, but public responses also place strong selective pressure on rapidly mutating pathogens to evade recognition by the most common TCRs.[5] Beyond pathogen responses, public TCRs have been associated with autoimmune conditions,[6,7] better health outcomes in B-cell lymphoma[8] and drug hypersensitivity reactions.[9] The origin of such shared or 'public' TCRs remains unclear, but both the formation and maintenance of the TCR repertoire contribute to the appearance of a public subset of TCRs.

While the potential diversity of the TCR repertoire is vast, biases in V(D)J recombination and selection contribute to observations of public TCRs. The α and β chains of each TCR are formed by recombining one each of V, D (for the β chain), and J gene segments from separate genomic loci. Unequal gene segment usage with correlations between V and J gene segment usage leads to more frequent TCRs that contain a particular V and J gene segment combination.[10] Gene segment usage biases are the strongest in inbred mice[11] and in monozygotic

Abbreviations: MHC, major histocompatibility complex; TCR, T-cell receptor twins.[12] Random insertion or deletion of nucleotides at the gene segment junctions contributes most of the diversity of the TCR repertoire.[13] Even with the random nucleotides at the gene segment junctions, however, sequences with fewer insertions or deletions and/or synonymous codon usage lead to a perceived generation bias for public sequences through convergent recombination.[14,15] Together, biased processes in V(D)J recombination lead to inferred generation probabilities of each unique TCR-β that vary by several orders of magnitude.[10]

Selective pressures on different TCRs further shape the frequency of TCRs in the repertoire. In the thymus, positive and negative selection through self-peptides presented by the MHC (self-pMHC) filters out non-reactive and highly self-reactive receptors from the developing repertoire.[16,17] Thymic selection thus increases the frequency of TCRs that have a slight affinity for self-pMHC.[18] Outside the thymus, peripheral selection similarly shapes the repertoire. The abundance of certain T-cell clonotypes may increase by homeostatic proliferation via interactions with self-pMHC[19] or receptor-independent mechanisms that increase survival and decrease death rates.[20,21]

Observational deep sequencing studies have found numerous public TCR-βs in unrelated individuals,[22,23] although the extent of sharing varies with sampling depth and cohort size.[10,14,15,24] The TCR-β monomer contributes significantly towards antigen recognition[25,26] and can identify unique TCR clones,[27,28] although the paired αβTCR can reveal antigen specificities that differ from the specificity inferred by the TCR-β alone.[27] Previous studies of the TCR repertoire have generally sequenced mRNA from antigen-experienced T cells as opposed to naive T cells and have not segregated by lineage (CD8 vs CD4), even though these subsets behave differently during immune surveillance and response.[29–31] While recent or past antigen exposure increases the frequency of clonally expanded T cells and the probability of observing shared receptor sequences, any shared antigen-experienced TCR must have started as a shared naive TCR. Further, the frequency of naive cells strongly influences the immune response upon pathogen exposure,[32] which means a correlation between naive TCR sharing and frequency could have an out-sized effect on the immune response. Studies of repertoires from bulk T cells can be confounded by the dynamics experienced by different T-cell subsets (CD4 vs. CD8),[33] which can be avoided by cell sorting prior to sequencing. Finally, mRNA sequencing data bias data towards productive TCR sequences, which reflect both recombination biases and selective forces; these forces can be separated by sequencing genomic DNA and examining non-productive TCR sequences, which do not encode functional receptors and do not experience selection.

Genomic DNA from T cells can be used to divide recombined sequences into 'productive' and 'non-productive' (i.e. frame shift or early stop codon) groups. These non-productive TCRs may be sequenced from T cells where a subsequent, independent rearrangement of the alternate chromosome resulted in a productive receptor. Selection operating on the productive receptor also indirectly drives the abundance of the non-productive rearrangement (if it exists) within the same T cell, but non-productive sequences as a whole should strictly reveal the effects of the generation process because non-productive sequences have an equal chance to be associated with a more favoured or less favoured productive TCR.[10,34] A further advantage of genomic DNA sequencing reads is their independence from TCR mRNA transcription levels, thus revealing the differential abundance of T cells.[35]

Here, we investigate how generation and selection processes contribute to a public repertoire by sequencing the CDR3β genomic region from sorted naive CD8+ T cells from three inbred mice. As these mice share the same MHC alleles, selection on self-pMHC during T-cell development acts equally in the three individual mice. Statistical models trained on non-productive sequences have been able to predict how many individuals share a particular amino acid TCR-β sequence,[24,36] but do not fully address abundances within a repertoire. While past work has focused on sharing based on TCR presence/absence, we take a more fine-scale view of diversity and also consider the frequencies of TCRs within each individual. We elucidate the effects of generation biases and TCR-mediated selection by comparing the TCR-β abundance distributions between productive (affected by both selection and generation biases) and non-productive (only affected by generation biases) subsets.

MATERIALS AND METHODS

Mice and cell sorting

We analysed the CDR3β region of naive T cells sampled from three C57BL/6 mice purchased from the Jackson laboratories (Bar Harbor, ME) and maintained under specific pathogen-free conditions at the ASU Biodesign Institute animal facilities. Mouse experiments were approved by the Institutional Animal Care and Use Committee at Arizona State University. We isolated CD8+ T cells from spleens of 6- to 8-week-old donor mice by positive immunomagnetic cell sorting (>95% CD8+, >95% CD44lo; Miltenyi Biotec) as previously described.[37] We prepared single-cell suspensions from splenocytes as previously described.[38] Briefly, we lysed erythrocytes with ammonium chloride lysis (ACK) buffer purchased from Lonza (Allendale, NJ) and performed FACS staining as previously described[39] in 96-well plates with fluorochrome-labelled monoclonal antibodies: anti-CD8 (clone 53-6.7), anti-CD44 (clone IM7) and anti-CD4 (clone GK1.5), purchased from BD Pharmingen (San Diego, CA) or eBioscience (San Diego, CA). Samples were then fixed in 1% paraformaldehyde solution, immediately acquired on a BD LSR II Fortessa flow cytometer (San Jose, CA) and analysed using FlowJo software (Tree-Star, Ashland, OR).

TCR sequencing and bioinformatic analysis

Three samples consisting of $1{\cdot}16 \times 10^6$, $1{\cdot}34 \times 10^6$ and $1{\cdot}48 \times 10^6$ CD8+ sorted cells from C57BL/6 mouse spleens were shipped to Adaptive Biotechnologies (Seattle, WA) for standard ImmunoSEQ TCR-β profiling.[40] Briefly, genomic DNA was amplified by multiplex PCR with equimolar pools of 45 forward V primers and 13 J reverse primers, covering the known functional murine TCR-β V and J gene segments. Additionally, each primer contains at the 5′ end the universal forward and reverse primer sequences compatible with the base Illumina sequencing technology. Sequencing was performed on a MiSeq analyser using 2 × 100 paired end reads, resulting in ~60 nucleotide-long reads where each read uniquely identifies a TCR-β sequence using the CDR3β region with V and J gene segment identities.[40] Read data were processed by ImmunoSEQ analyser for correction of PCR biases informed by synthetic repertoires,[41] counting read abundances,[42] and identification of germline gene segments, the number of insertions and deletions and functional status with standardized definitions from the international ImMunoGeneTics information system.[43] We defined a nucleotide TCR-β by the full sequence read, and the amino acid TCR-β by the V and J gene segment identities with the translated CDR3β region of the read.

Model testing

We calculated empirical cumulative distribution functions (ECDFs) of the unique TCR abundances from a pooled data set of all three mice, dividing the repertoire by sharing level (private to the individual, shared among two individuals, shared among all three individuals) and functional status (productive vs. non-productive). In this application of ECDFs, similar to species-abundance distributions,[44] we treat each unique TCR-β as a separate species and examine the abundance distribution. Specifically, the x-axis denotes the abundance of TCR-βs, and the y-axis denotes the cumulative fraction of unique TCR-βs with abundance equal to or less than the x-axis value. A repertoire containing very low-abundance TCR-βs results in a right-shifted ECDF, while a repertoire dominated by high abundance results in a left-shifted function. We assessed the uncertainty of ECDFs in different subsets of sequences with 100 bootstrap replicates of the repertoire. For each replicate, we sampled the same number of total TCR-βs from each mouse (equal to the smallest number of total reads per mouse in the original data), recalculating abundance distributions of unique TCR-βs in each set of bootstrap replicates. For unique TCR-βs shared across individuals in the pooled data set, we randomly chose one TCR-β abundance (of the abundances for TCR-β found in the two or three mice) to avoid the reduction in variance that would be caused by using the mean or median of the sequence abundance in multiple mice. Bootstrapped ECDFs of the individual mice are shown in Figure S1. Because the productive subset was more deeply sampled compared with the non-productive subset, we also performed the same analysis with down-sampled productive sequences.

We performed non-parametric tests to estimate the statistical significance of the correlation between abundances of shared TCR-βs and dissimilarity in gene segment usages between subsets of the sampled repertoires. For abundance correlations of pairwise shared TCR-β, we permuted the abundances of TCR-βs in one of each pair of mice, recalculating the Pearson correlation coefficient 1000 times. For gene segment usage, we compared subsets of TCR-β based on binary sharing (private vs. shared) and functional status (productive vs. non-productive). We compared the V and J gene segment usage proportion of unique TCR-βs across each subset, calculating the sum of squared deviation in gene usage proportions between each subset. We estimated 95% confidence intervals of the sum of squared deviation in gene usage by 100 bootstrap replicates of individual repertoires, visualizing the gene usage differences as a heat map.

In addition to the larger sample size of productive sequences compared with non-productive sequences, we also considered the broader length distributions of non-productive sequences compared with productive sequences. By definition, the TCR-β lengths of productive sequences are multiples of three base pairs (e.g. 27, 30, 33), while the lengths of non-productive sequences are not restricted in the same way (e.g. 27, 28, 29, 30, 31). Therefore, non-productive sequences as a whole have more opportunities for sharing. When calculating sharing proportions, we correct for this length effect by selecting non-productive sequence lengths with the most sequences within one base pair of the corresponding productive sequence length (e.g. 26 in non-productive and 27 in productive).

Calculating the public fraction of a repertoire

We calculate the estimated public fraction, defined as the fraction of the repertoire with size N that contains sequences with a generation probability greater than 1/N, according to equation 11 in Ref. 45. We approximated the integral over the density of sequence generation probability by the rectangle approximation, taking kernel density estimates over 2^{15} equally spaced points for the generation probability distribution of unique sequences shown in Figure 4. We calculated the public fraction over a range of repertoire sizes for generation probability distributions of different subsets of the repertoire, including productive-only and non-productive-only.

Estimating TCR generation probability with IGoR

We applied IGoR[45] to estimate generation probabilities of TCRs with a recombination model inferred from the non-productive sequences of our data. For the IGoR alignment step, we supplied the set of germline IMGT V and J genes for genomic templates,[43] setting V and J gene offset bounds for where the genomic templates may align given the short length of our reads (V offset: −350, −150; J offset: 0, 80). With default parameters, we inferred a recombination model given the non-productive sequences of each mouse individually and together, and evaluated the generation probability of the observed TCR sequences.

RESULTS

Throughout our analysis, we compare sequences in the productive and non-productive naive repertoires to reveal the effects of thymic and peripheral selection beyond the biases in the generation process. We begin by examining the unique sequences in a pooled data set from all three mice and move on to analysing abundances.

Productive TCR-β sequences are shared more than non-productive sequences

If we define a shared TCR-β as one that is found at any frequency in at least two of our three mice, then 3·3% of the unique productive nucleotide sequences were shared, while only 0·65% of the unique non-productive sequences were shared (see Table 1). We consider that, exclusive of other factors, frame shifts lead to an approximate 1:2 ratio of the number of possible productive sequences to possible non-productive sequences, reducing the probability of sharing in the non-productive repertoire. Therefore, we might expect twofold less sharing in the non-productive repertoire. In addition, deeper sampling of productive relative to non-productive sequences further contributes to the shortfall in the non-productive repertoire. To account for the latter factor, we down-sampled productive sequences from each individual to the size of each individual non-productive repertoire and recalculated sharing proportions for 1000 bootstrap replicates. From this procedure, we still saw a higher sharing proportion of 3·23% in the productive repertoire, with a 95% bootstrap confidence interval for the difference between productive and non-productive sharing proportions of [2·60%, 2·62%], which remains well more than twofold the sharing proportion in the non-productive repertoire (0·65%).

Table 1. Number of unique TCR-β sequences

Sequence level	Sequence status	Private	Two-way	Three-way
Nucleotide	Productive	311 524	9221	1253
Nucleotide	Non-productive	139 672	857	56
Amino acid	Productive	211 765	17 959	4808
Amino acid	Non-productive	6115	161	16

A more subtle difference in TCR-β length distribution between productive and non-productive subsets may also affect sharing proportions. Productive sequences are constrained to have TCR-β lengths in multiples of three, while non-productive sequences include out-of-frame sequences. Because non-productive sequences have more length categories, comparisons between productive and non-productive sequences should be restricted to those with similar lengths. We thus down-sampled the productive subset and selected non-productive TCR-βs with similar lengths from each mouse as described in Methods. When we calculate the sharing proportion in 1000 bootstrap samples, we still observe a higher sharing proportion in productive sequences compared with non-productive sequences (2·49% [2·47, 2·51% 95% CI of difference between productive and non-productive sharing proportion]).

This pattern also holds when we translate the TCR-β nucleotide sequences into amino acids and consider amino acid clonotypes composed of the V gene segment, J gene segment and CDR3β amino acid sequence. Codon degeneracy reduces the number of total unique clonotypes in both the productive and non-productive repertoires. Omitting out-of-frame rearrangements that cannot be meaningfully translated further reduces the non-productive repertoire sample size. Even after down-sampling productive amino acid TCR-βs, the proportion of unique shared sequences in the productive repertoire continues to be higher than in the non-productive repertoire (mean difference of 1·79% [1·59%, 1·99% 95% CI]).

Finally, while the results described above derived from pooling all three mice together, we found this same pattern of higher sharing proportions in the productive compared with non-productive repertoires in each mouse individually (results not shown).

Shared TCRs correlate in frequency between mice

We next sought to consider not just TCR presence/absence in sharing but also relative abundances of unique TCR sequences. Given that clonal expansion during an immune response occurs as an exponential growth process,[46] we compare abundances on a logarithmic scale. Specifically, pairwise comparisons of log-transformed abundances of shared productive TCR-βs reveal statistically significant correlation coefficients for all pairs of mice (see Figure 1A–C with Pearson's $r = 0{\cdot}36$, 0·43, 0·33, $P < 0{\cdot}001$ by permutation test). Although there were fewer non-productive sequences, the same analysis applied to non-productive TCR-β abundances yielded higher correlation coefficients of 0·46, 0·52, and 0·41 ($P < 0{\cdot}001$ by permutation test) (see Figure 1D–F).

Correlations of abundance measures show that consistent, reproducible processes drive the abundances of shared naive TCR-βs of different individuals. As non-productive sequences are not translated into functional receptors, their observed abundances result from repeated generation in distinct T cells with the dynamics mediated by the productively rearranged receptor in the same T cell. A particular non-productive rearrangement may rise in abundance if its partner productive rearrangement promotes clonal expansion or survival. However, the independence of V(D)J recombination between the two chromosomes means the non-productive subset as a whole represents neutral sequences with respect to thymic and peripheral selection. As the abundances of sequences in the non-productive subset result from selectively neutral processes relative to the productive subset, the higher correlation of non-productive sequences illustrates how consistent generation biases across individuals contribute to TCR-β sharing. While consistent receptor generation processes may explain the higher correlations in abundances of the non-productive shared sequences, between-individual differences in selective pressures may then alter the correlation of abundances of shared productive sequences.

Shared TCR-βs have higher frequencies than private TCR-βs, but only for productive sequences

When we compare the frequency distributions of shared to private TCR-βs, we observe a striking difference between productive and non-productive sequences. The empirical cumulative distribution function (ECDF) curves show the distribution of TCR-β abundance frequencies of each unique TCR-β sequence, where a sample with TCR-βs found mostly at smaller abundances would have a right-shifted ECDF relative to a sample with TCR-βs found at higher abundances. In the productive repertoire, the frequency distribution of private sequences is shifted significantly lower than two-way shared sequences, which in turn is shifted significantly lower than three-way

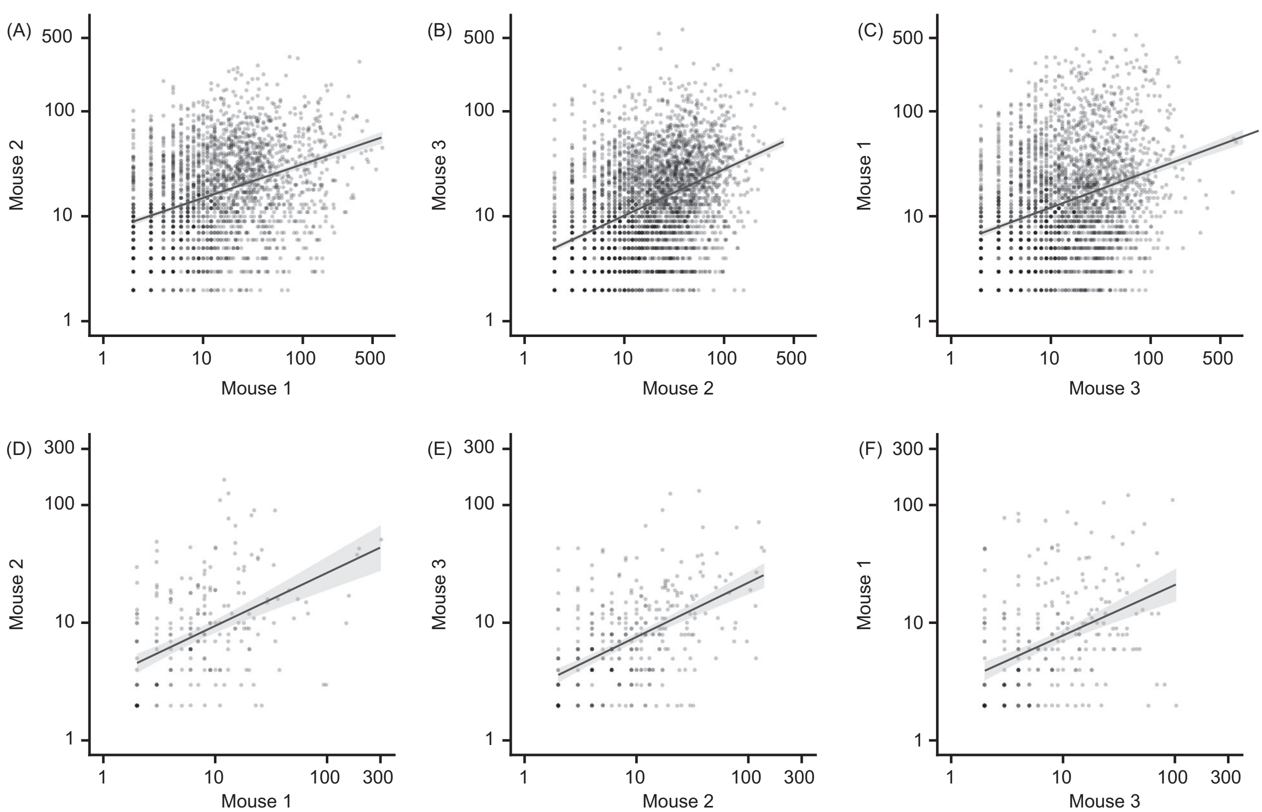

Figure 1. Abundances of shared TCR-βs are positively correlated in pairwise comparisons of samples from different mice. (A–C) Log abundances of productive sequences for different pairs of mice with Pearson's $r = 0{\cdot}36$, 0·43 and 0·33 ($P < 0{\cdot}001$), respectively, are shown. (D–F) The same pairwise comparisons for non-productive sequences with Pearson's $r = 0{\cdot}46$, 0·52 and 0·41 ($P < 0{\cdot}001$), respectively, are shown. We estimate P-values by 1000 permutations of TCR-β abundances. We plot a linear regression line with 95% confidence intervals

shared sequences (Figure 2). In the non-productive repertoire, we observe the *opposite* pattern with the frequency distribution of private sequences being higher than two-way shared sequences, which in turn is higher than three-way shared sequences. These patterns also hold with amino acid clonotypes (see Figure S2), although the result is not significant for non-productive translated sequences, in part due to small sample sizes.

We further confirmed that these results were not driven by sequencing error-prone rare T cells or contamination from high-abundance memory T cells. As approximately 50% of all unique sequences were found in copy numbers of 10 or less, we analysed the subset of data that contain sequences with read abundances greater than 10 (Figure S3B). We observed the same shifts in distributions in the productive repertoire without low-abundance sequences, indicating the expansion of specific T cells carrying certain TCR-β sequences drove the differences. Separately, antigen-experienced memory TCRs are expected to be found at high abundances, some of which may contaminate the naive TCR data due to imperfect sorting. To control for this possibility, we analysed the subset of data that contain sequences below the 95th percentile abundance of TCR-βs in each individual. Again, we observed the same shifts in distribution in the productive repertoire towards higher abundance as sharing increases (Figure S3A). We performed both comparisons while down-sampling the productive repertoire to the size of the non-productive repertoire, to reduce the effect of deeper sampling of productive sequences.

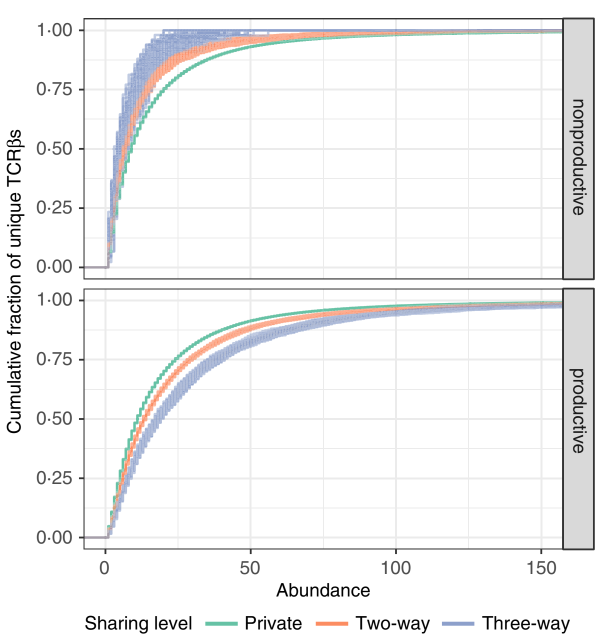

Figure 2. Empirical cumulative distribution of nucleotide read abundances, with 100 bootstrapped replicates. The distribution of abundances shifts towards higher abundances as sharing increases in the productive, but not in the non-productive, repertoire

Possible explanations for the frequency difference in public sequences

Given this difference in shared TCR-βs between productive and non-productive subsets, we looked for systematic differences between the two repertoires that might explain the increased abundances of shared sequences in the productive repertoire. Increased abundances could be driven by the same factor as increased sharing—namely increased rates of generation. Factors potentially affecting a TCR-β generation rate include the number of added nucleotides at V(D)J junctions, gene segment usage and codon degeneracy. We compared subsets of the repertoire and found different aspects of the generation process that impact sharing rates, but do not fully account for abundance differences between shared and private TCR-βs of the productive and non-productive repertoires.

In general, shorter TCR-β sequences are more likely to be shared across multiple individuals because they have fewer non-templated inserted nucleotides to be matched at the V-D and D-J junctions by the process of convergent recombination.[14,15] The higher rates of convergent recombination of shorter TCR-βs increase their frequencies in the repertoire. Unique productive and non-productive sequences in general had similar lengths, although non-productive were marginally longer. Within both subsets, shared sequences were shorter than private sequences, from a mean of 37 (private) to 35·4 (shared) nucleotides in non-productive and 36 to 34 nucleotides in productive sequences (Figure 3). The decrease in size for productive sequences was significantly greater than for non-productive ($P < 0{\cdot}001$, *t*-test for mean difference in productive is equal to mean difference in non-productive), although the magnitude of the effect was small.

Biased gene segment usage itself can make certain sequences more frequent.[47] We examined V gene segment usage as a function of degree of sharing in both repertoires to determine how recombination biases may influence public sequences, specifically whether gene usage frequencies differed between public and private TCRs. We measured the sum of squared deviation of gene usages across sequences within the same functional status but across different sharing levels and vice versa over 1000 bootstrap replicates. We found the greatest gene usage difference in the shared sequences between productive and non-productive subsets, while the magnitude of difference was much smaller within productive or non-productive sequences (see Figure S5A and B depicting mean deviations). We observed similar differences for the J gene segment usage, but with larger magnitudes of squared deviations. These results show that gene usages were very similar across subsets of the repertoire, except for

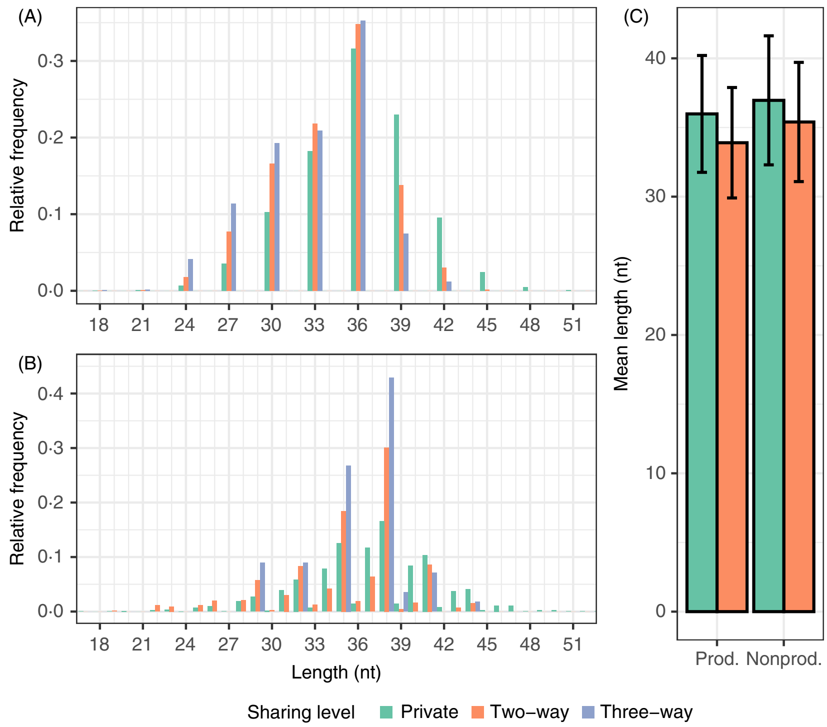

Figure 3. CDR3 length distribution of unique TCR-βs within subsets defined by functional status and sharing level across three mice. Nucleotide CDR3 lengths showed similar ranges (productive: 12 to 63, non-productive: 11 to 65), with shifts across sharing levels in both productive (A) and non-productive (B) subsets. (C) Differences between the mean CDR3 length of private and shared sequences in both the productive (private: 36 [4·2 SD], shared: 33·9 [4·0 SD]) and non-productive (private: 37 [4·7 SD], shared: 35·4 [4·3 SD]) subsets were small, but statistically significant (Welch's *t*-test, $P < 0{\cdot}001$)

between shared productive and shared non-productive subsets.

In addition to gene segment usage and TCR-β length, other more subtle factors can influence the generation probability[10,24] and thus the sharing probability of TCRs. We applied IGoR[45] to estimate the generation probabilities of both productive and non-productive sequences, fitting the model to the non-productive sequences in our samples. IGoR predicts private sequences to have a lower inferred generation probability than public sequences (Figure 4), although inferred generation probability did not correlate with sequence abundances (Figure S4A; productive adjusted $r^2 = 0{\cdot}000025$, non-productive adjusted $r^2 = 0{\cdot}0063$). Additionally, we found that when binned by inferred generation probabilities such that the total number of unique TCR-βs was the same in each bin (~2600 per non-productive bin, ~6200 per productive bin), sequences were more likely to be shared as generation probability increased (Figure S4B). Generation probabilities correlate well with the sharing of unique sequences in both the productive and non-productive repertoire, but cannot fully explain the increase in frequency of shared productive sequences.

While our observations of shifts in the frequency distribution of TCR abundance focused on nucleotide sequences, we also evaluated how convergent recombination at the amino acid level affected sharing through codon degeneracy.[14,15] Analysis of unique non-productive amino acid TCR-βs was limited due to a smaller number of in-frame TCR-β nucleotide sequences. We found that for amino acid TCR-βs, 97% of shared, unique productive sequences were encoded by more than one nucleotide sequence, while only 93·3% of shared, unique non-productive sequences were multiply-encoded. In both productive and non-productive repertoires, the mode of the number of degenerate nucleotide sequences per amino acid TCR-β increases, from only one nucleotide sequence per amino acid sequence in the private subset to three nucleotide sequences per amino acid sequences in three-way shared sequences. Further, three-way shared productive amino acid TCR-βs may be encoded by up to 20 unique nucleotide sequences (see Figure S2B). This summary, however, is confounded by the fact that a longer amino acid TCR-β has more degenerated nucleotide sequences and we have many more productive than non-productive sequences in our sample. To assess the statistical significance of these differences, we down-sampled the number of productive sequences to the number of non-productive sequences while also limiting sequences to the most common lengths of 35–37 nucleotides long. In this

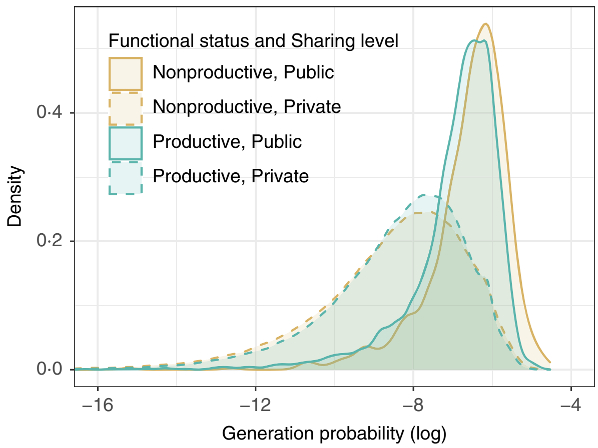

Figure 4. Distribution of log generation probabilities of unique sequences by sharing level and functional status inferred by IGoR. Generation probability inferences failed for approximately 5% of the unique sequences in each functional subset. Shared TCR-βs in both non-productive (860 sequences) and productive (10 144 sequences) subsets show a narrower distribution of generation probabilities centred on a higher value. Private TCR-βs in the productive (297 845 sequences) and non-productive (132 644 sequences) show similarly broad distributions of generation probabilities centred two magnitudes lower than the shared TCR-βs

analysis, we did not observe a difference in the distribution of multiply-encoded sequences in the productive and non-productive repertoire. Down-sampling excluded most of the TCR-βs that were encoded by more than four nucleotide sequences and led to a significant shift in the mode of multiply-encoded TCR-βs. In general, productive TCR-βs were more degenerate, and shared TCR-βs also showed a higher mean degeneracy compared with private sequences, which contribute to higher abundances. These results for amino acid TCR-βs do not explain the observed shifts in nucleotide sequence abundance distributions.

DISCUSSION

We found that unique shared TCR-βs were over-represented in the productive repertoire of naive CD8+ T cells in three inbred mice. The distribution of TCR-β abundances revealed this pattern even more strongly with shared sequences having higher abundances in the productive, but not in the non-productive, repertoire. As the primary difference between productive and non-productive sequences is that the former are expressed in protein form, these shared productive receptors presumably increased in abundance through thymic and peripheral intra-individual selection pressures, in addition to the shared generation biases common between the three mice.

Alternatively, differences between the private and shared TCR-β sequence abundance distributions could arise from only generation biases or technical artefacts with no contribution from intra-individual selection. We control for generation biases by directly comparing the productive and non-productive repertoires in the same individual. As both repertoires are generated via the same process, generation biases should be common as well, and thus, comparisons between productive and non-productive repertoires should remove any bias and reveal the contribution of selection. Similarly, while technical artefacts such as PCR amplification bias may increase the apparent abundance of some sequences,[41,48] such bias should again affect both repertoires given that both have similar V/J segment usage (see Figure S5). In an effort to minimize the effect of PCR bias, steps in post-sequence processing included adjustments for sequence abundances based on V or J segment usage,[40] and previous work demonstrated the repeatability and sensitivity of the sequencing and quantification assay in detecting specific TCR-βs from low-abundance clones.[42] Another possible technical artefact could arise from sequencing errors that generate a large number of rare and false receptor sequences. We controlled for this possibility by performing our analysis restricted to only higher abundance sequences and found the same qualitative results (see Figure S3B). Contamination from memory T cells may also influence shared TCR abundance distributions, as memory T cells are expected to be found at high abundance and may then share TCR sequences due to exposure to shared foreign antigens. We tested for the effect of memory contamination by excluding high-abundance reads and recalculating ECDFs, and again found similar trends of increased TCR abundance as sharing increases in the productive but not in the non-productive subset.

While selection is necessary to explain the relative differences between productive and non-productive repertoires, prior research has shown that convergent recombination plays a key role in driving shared TCR-β sequences on an absolute scale.[14] Indeed, the formation of public sequences can be attributed to biases in gene segment usage and shorter TCR-β lengths that allow easier convergent recombination and thus a higher generation probability,[49,50] particularly between genetically identical individuals.[12] Given this expectation, we examined our data for TCR-β length and gene segment usage biases.

First, we compared the TCR-β length distributions in shared and unshared repertoires and found that shared sequences were shorter in both the productive and non-productive repertoires. The productive sequences were on average one nucleotide shorter than non-productive, which contributes towards more sharing in the productive repertoire as a whole compared with the non-productive repertoire. At the same time, non-productive sequences could be found at more length categories than productive sequences, which may inflate sharing in the non-productive repertoire as a whole. We corrected for differences in length distribution and sample size between productive and non-productive repertoires and found a higher sharing proportion in the productive repertoire compared with the non-productive repertoire, showing that functional status drove total sharing probability. Second, we compared gene segment usages between the four categories of the repertoire. We expected gene usage differences between shared and unshared sequences to be small in the non-productive subset, because their presence should only be driven by a common generation processes. We found that the largest gene segment usage differences were between productive and non-productive sequences, regardless of the sharing levels (Figure S5). The gene usage differences between shared and unshared sequences in the productive repertoire were the smallest of all the comparisons, showing that even if gene segment usage influences the functional status of a sequence, it does not have a large effect on whether a sequence is public or private within the productive repertoire. These features of length and gene usage were insufficient to predict whether a TCR-β was public or private with a machine learning classifier.[51]

In addition to examining the directly observable characteristics of our data, we also applied a probabilistic generative model of VDJ recombination using the software IGoR, which captures more subtle aspects of the recombination process, to estimate TCR-β generation probabilities.[10,45] As the IGoR model does not use the observed abundances, this approach provided independent evidence for the action of selection. Although the inferred probabilities of our sequence data were not predictive of sequence abundance in the non-productive or productive repertoires (Figure S4A), we found that the proportion of shared sequences increases with higher quantiles of generation probability (Figure S4B). We also observed several TCR-βs with relatively high abundances yet relatively low generation probabilities (Figure S4A). These generation probability estimates were robust to whether recombination model parameters were inferred from the non-productive sequences of individual mice or all mice together. These results suggest minor biases inside the IGoR estimation process or the abundance estimation process. We also considered how shifts in the distribution of generation probabilities (Figure 4) influence estimates of the fraction of 'public' sequences, defined as sequences that have generation probabilities greater than the reciprocal of the total number of sequences.[24] Using bootstrapped samples of generation probability distributions from different subsets of the repertoire illustrated in Figure 4, we find that the highest public fraction came from repertoires with generation probability distributions similar to the empirically public sequences (Figure S6A), but that the productive subset as a whole shows only a small, but statistically significant, increase in public fraction compared with the non-productive subset (Figure S6B).

These results show that large shifts in the generation probability distribution lead to increases in sharing proportions, but do not fully account for opposite trends of the abundance distributions (Figure 2) nor the difference in observed sharing proportions between productive and non-productive subsets. Indeed, introducing a sequence-independent selection coefficient improved the predictions of sharing numbers in empirical samples.[24] Abundance distribution shifts of TCR-βs in the productive subset arise from both biases in generation and biases in proliferation or survival, even for naive T cells.[28,52] We cannot distinguish between the mechanisms of enhanced proliferation or survival of shared T cells, which could arise through several mechanisms.[53] As young mice have relatively short-lived T cells and strong thymic output,[54] preferential selection for attributes inherent to a new T cell may afford a relative fitness advantage before T-cell adaptation dominates.[53] This preferential selection likely involves self-antigens,[55] which in turn suggests that shared sequences may play a role in autoimmunity. Despite their association with pathology, receptors with self-reactivity may be important to clearing infections, analogous to autoimmune-recognizing antibodies,[56] where the strength of self-recognition also correlates with reactivity towards foreign antigens.[57] Experiments have shown how the relative strength of self-pMHC-TCR interactions drives proliferation of adoptively transferred naive CD8+ T cells, which compete among T cells bearing different receptors.[58] Besides antigen recognition, differences in peptide availability and/or MHC expression may also contribute to differential selective pressures.[59,60] Experimental work in humanized mouse models found when compared to unshared TCR-βs, public TCR-βs were enriched for type 1 diabetes-associated TCRs and higher cross-reactivity to different MHC alleles, but potentially weaker interaction with self-peptides (allowing for escape from negative selection).[61]

We conclude that selection contributes towards sharing between individuals, given the generation of a particular sequence, but do not quantify the relative contribution between different forms of selection and generation biases. Although shared immune history drives receptor sharing in the repertoire as a whole,[25,62] by restricting our analysis to sorted naive T cells we exclude selection due to experience with common foreign antigens. For naive T cells, only TCR-mediated interactions are expected to give certain T cells a selective advantage, as advantages from non-TCR effects (such as cytokine sensitivity) should be uncorrelated with a specific TCR-β. Our use of the TCR-β abundance information strengthened our conclusions and revealed dynamics in the naive T-cell repertoire that would be largely hidden by examining only unique sequences. Paired αβTCR sequencing[63–66] would provide a more complete picture of the public repertoire, although shared αβTCR may be exceedingly rare except for epitope-specific responses.[67,68] Further, few non-productive α chains may be sampled due to the potential for multiple rescue attempts of a non-productive TCR-α locus rearrangement.[69] Recent work identified 26 paired αβTCRs shared between any two of five individuals from high-throughput single-cell sequencing of CD4+ and CD8+ T cells, which shows biases in generation and pairing of the individual chains in addition to convergent selection of non-naive T cells.[71] While technical biases pose a challenge to using abundance data, comparison of productive and non-productive repertoires provides a means to overcome these biases and reveals how within-individual selection contributes to TCR-β sharing.

ACKNOWLEDGMENTS

HHY was supported by funding from the US Department of Education GAANN Program. LNS was supported in part by NIH R43CA236142. JNB was supported by NIH R21 AI125827 and R21 CA196460. PLFJ was supported in part by NIH R00 GM104158. LS, MG and JB conducted mouse experiments and performed sequencing. HHY performed analyses. HHY and PJ designed analyses and wrote the manuscript. LS, JB, HHY and PJ edited the manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

TCR sequence data are available upon request from the corresponding author.

REFERENCES

1 Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature* 1988; **334**:395.

2 Doherty PC, Riberdy JM, Belz GT. Quantitative analysis of the CD8+ T-cell response to readily eliminated and persistent viruses. *Philos Trans R Soc B Biol Sci.* 2000; **355**:1093–101.

3 Casrouge A, Beaudoing E, Dalle S, Pannetier C, Kanellopoulos J, Kourilsky P. Size estimate of the TCR repertoire of Naive mouse splenocytes. *J Immunol.* 2000; **164**:5782–7.

4 Li H, Ye C, Ji G, Han J. Determinants of public T cell responses. *Cell Res.* 2012; **22**:33–42.

5 Miles JJ, Douek DC, Price DA. Bias in the A β T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol Cell Biol.* 2011; **89**:375–87.

6 Zhao Y, Nguyen P, Ma J, Wu T, Jones LL, Pei D, *et al.* Preferential use of public TCR during autoimmune encephalomyelitis. *J Immunol.* 2016; **196**:4905–14.

7 Miles JJ, Bulek AM, Cole DK, Gostick E, Schauenburg AJA, Dolton G, *et al.* Genetic and structural basis for selection of a ubiquitous T cell receptor deployed in Epstein-Barr virus infection. *PLoS Pathog.* 2010; **6**:e1001198.

8 Keane C, Gould C, Jones K, Hamm D, Talaulikar D, Ellis J, *et al.* The T-cell receptor repertoire influences the tumor microenvironment and is associated with survival in aggressive B-cell lymphoma. *Clin Cancer Res.* 2017; **23**:1820–8.

9 Ko T-M, Chung W-H, Wei C-Y, Shih H-Y, Chen J-K, Lin C-H, *et al.* Shared and restricted T-cell receptor use is crucial for Carbamazepine-Induced Stevens-Johnson Syndrome. *J Allergy Clin Immunol.* 2011; **128**:1266–76.e11.

10 Murugan A, Mora T, Walczak AM, Callan CG. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci USA.* 2012; **109**:16161–6.

11 Rettig TA, Ward C, Bye BA, Pecaut MJ, Chapes SK. Characterization of the Naive Murine antibody repertoire using unamplified high-throughput sequencing. *PLoS One* 2018; **13**:e0190982.

12 Rubelt F, Bolen CR, McGuire HM, Heiden JAV, Gadala-Maria D, Levin M, *et al.* Individual heritable differences result in unique cell lymphocyte receptor repertoires of Naïve and antigen-experienced cells. *Nat Commun.* 2016; **7**:11112.

13 Rossjohn J, Gras S, Miles JJ, Turner SJ, Godfrey DI, McCluskey J. T cell antigen receptor recognition of antigen-presenting molecules. *Annu Rev Immunol.* 2015; **33**:169–200.

14 Venturi V, Kedzierska K, Price DA, Doherty PC, Douek DC, Turner SJ, *et al.* Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc Natl Acad Sci USA.* 2006; **103**:18691–6.

15 Quigley MF, Greenaway HY, Venturi V, Lindsay R, Quinn KM, Seder RA, *et al.* Convergent recombination shapes the clonotypic landscape of the Naive T-Cell repertoire. *Proc Natl Acad Sci USA.* 2010; **107**:19414–9.

16 Klein L, Kyewski B, Allen PM, Hogquist KA. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nat Rev Immunol.* 2014; **14**:377–91.

17 Yates AJ. Theories and quantification of thymic selection. *Front Immunol.* 2014; **5**:13.

18 Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W, *et al.* T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res.* 2014; **24**:1603–12.

19 Surh CD, Sprent J. Homeostasis of Naive and memory T cells. *Immunity* 2008; **29**:848–62.

20 Clarke SRM, Rudensky AY. Survival and homeostatic proliferation of naive peripheral CD4+ T cells in the absence of self peptide: MHC complexes. *J Immunol.* 2000; **165**:2458–64.

21 Johnson PLF, Yates AJ, Goronzy JJ, Antia R. Peripheral selection rather than thymic involution explains sudden contraction in Naive CD4 T-cell diversity with age. *Proc Natl Acad Sci USA.* 2012; **109**:21432–7.

22 Venturi V, Price DA, Douek DC, Davenport MP. The molecular basis for public T-cell responses? *Nat Rev Immunol.* 2008; **8**:231–8.

23 Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, *et al.* Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci Transl Med.* 2010; **2**:47ra64.

24 Elhanati Y, Sethna Z, Callan CG, Mora T, Walczak AM. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunol Rev.* 2018; **284**:167–79.

25 Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* 2017; **547**:94–8.

26 Jorgensen JL, Esser U, Fazekas de St. Groth B, Reay PA, Davis MM. Mapping T-cell receptor–Peptide contacts by variant peptide immunization of single-chain transgenics. *Nature* 1992; **355**:224–30.

27 Carter JA, Preall JB, Grigaityte K, Goldfless SJ, Jeffery E, Briggs AW, *et al.* Single T cell sequencing demonstrates the functional role of αβ TCR pairing in cell lineage and antigen specificity. *Front Immunol.* 2019; **10**:1516.

28 de Greef PC, Oakes T, Gerritsen B, Ismail M, Heather JM, Hermsen R, *et al.* The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes. *eLife* 2020; **9**:e49900.

29 De Boer RJ, Homann D, Perelson AS. Different dynamics of CD4 $^{+}$ and CD8 $^{+}$ T cell responses during and after acute lymphocytic choriomeningitis virus infection. *J Immunol.* 2003; **171**:3928–35.

30 Kaur A, Hale CL, Ramanujan S, Jain RK, Johnson RP. Differential dynamics of CD4+ and CD8+ T-lymphocyte proliferation and activation in acute simian immunodeficiency virus infection. *J Virol.* 2000; **74**:8413–24.

31 Mandl JN, Liou R, Klauschen F, Vrisekoop N, Monteiro JP, Yates AJ, *et al.* Quantification of lymph node transit times reveals differences in antigen surveillance strategies of naive CD4+ and CD8+ T cells. *Proc Natl Acad Sci USA.* 2012; **109**:18036–41.

32 Kotturi MF, Scott I, Wolfe T, Peters B, Sidney J, Cheroutre H, *et al.* Naive precursor frequencies and MHC binding rather than the degree of epitope diversity shape CD8+ T cell immunodominance. *J Immunol.* 2008; **181**:2124–33.

33 Emerson R, Sherwood A, Desmarais C, Malhotra S, Phippard D, Robins H. Estimating the ratio of CD4+ to CD8+ T cells using high-throughput sequence data. *J Immunol Methods.* 2013; **391**:14–21.

34 Manfras BJ, Terjung D, Boehm BO. Non-productive human TCR β chain genes represent V-D-J diversity before selection upon function: insight into biased usage of TCRBD and TCRBJ genes and diversity of CDR3 region length. *Hum Immunol.* 1999; **60**:1090–100.

35 Heather JM, Ismail M, Oakes T, Chain B. High-throughput sequencing of the T-Cell receptor repertoire: pitfalls and opportunities. *Brief Bioinform.* 2017;**19**:554–65.
36 Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* 2019; **566**:393–7.
37 Mora JR, Bono MR, Manjunath N, Weninger W, Cavanagh LL, Rosemblatt M, *et al.* Selective imprinting of gut-homing T cells by Peyer's patch dendritic cells. *Nature* 2003; **424**:88–93.
38 Murali-Krishna K, Altman JD, Suresh M, Sourdive DJD, Zajac AJ, Miller JD, *et al.* Counting antigen-specific CD8 T cells: a reevaluation of bystander activation during viral infection. *Immunity* 1998; **8**:177–87.
39 Altman JD, Moss PAH, Goulder PJR, Barouch DH, McHeyzer-Williams MG, Bell JI, *et al.* Phenotypic analysis of antigen-specific T lymphocytes. *Science* 1996; **274**: 94–6.
40 Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, *et al.* Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 2009; **114**:4099–107.
41 Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung M-W, Parsons JM, *et al.* Using synthetic templates to design an unbiased multiplex PCR assay. *Nature Commun.* 2013; **4**:3680.
42 Robins H, Desmarais C, Matthis J, Livingston R, Andriesen J, Reijonen H, *et al.* Ultrasensitive detection of rare T cell clones. *J Immunol Methods.* 2012; **375**:14–9.
43 Monod MY, Giudicelli V, Chaume D, Lefranc M-P. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics* 2004; **20**:i379–i385.
44 McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK, *et al.* Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol Lett.* 2007; **10**:995–1015.
45 Marcou Q, Mora T, Walczak AM. High-throughput immune repertoire analysis with IGoR. *Nature Commun.* 2018; **9**:561.
46 Burnet F. A modification of Jerne's theory of antibody production using the concept of clonal selection. *Austr J Sci.* 1957; **20**:119–21.
47 Ndifon W, Gal H, Shifrut E, Aharoni R, Yissachar N, Waysbort N, *et al.* Chromatin conformation governs T-cell receptor J gene segment usage. *Proc Natl Acad Sci USA.* 2012; **109**:15865–70.
48 Sint D, Raso L, Traugott M. Advances in multiplex PCR: balancing primer efficiencies and improving detection success. *Methods Ecol Evol.* 2012; **3**:898–905.
49 Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T, *et al.* A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J Immunol.* 2011; **186**:4285–94.
50 Li H, Ye C, Ji G, Wu X, Xiang Z, Li Y, *et al.* Recombinatorial biases and convergent recombination determine interindividual TCR sharing in murine thymocytes. *J Immunol.* 2012; **189**:2404–13.
51 Greiff V, Weber CR, Palme J, Bodenhofer U, Miho E, Menzel U, *et al.* Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. *J Immunol.* 2017; **199**:2985–97.
52 Hogan T, Gossel G, Yates AJ, Seddon B. Temporal fate mapping reveals age-linked heterogeneity in naive T lymphocytes in mice. *Proc Natl Acad Sci USA.* 2015; **112**: E6917–E6926.
53 Rane S, Hogan T, Seddon B, Yates AJ. Age is not just a number: Naive T cells increase their ability to persist in the circulation over time. *PLoS Biol.* 2018; **16**:e2003949.
54 den Braber I, Mugwagwa T, Vrisekoop N, Westera L, Mögling R, Bregje de Boer A, *et al.* Maintenance of peripheral Naive t cells is sustained by thymus output in mice but not humans. *Immunity* 2012; **36**:288–97.
55 Correia-Neves M, Waltzinger C, Mathis D, Benoist C. The shaping of the T cell repertoire. *Immunity* 2001; **14**:21–32.
56 Rivera-Correa J, Rodriguez A. Divergent roles of antiself antibodies during infection. *Trends Immunol.* 2018; **39**:515–22.
57 Mandl JN, Monteiro JP, Vrisekoop N, Germain RN. T cell-positive selection uses self-ligand binding strength to optimize repertoire recognition of foreign antigens. *Immunity* 2013; **38**:263–74.
58 Ge Q, Bai A, Jones B, Eisen HN, Chen J. Competition for self-peptide-MHC complexes and cytokines between Naive and Memory CD8+ T cells expressing the same or different T cell receptors. *Proc Natl Acad Sci USA.* 2004; **101**:3041–6.
59 Neefjes JJ, Ploegh HL. Allele and locus-specific differences in cell surface expression and the association of HLA class I heavy chain with β2-microglobulin: differential effects of inhibition of glycosylation on class I subunit association. *Eur J Immunol.* 1988; **18**:801–10.
60 Greene JM, Wiseman RW, Lank SM, Bimber BN, Karl JA, Burwitz BJ, *et al.* Differential MHC class I expression in distinct leukocyte subsets. *BMC Immunol.* 2011; **12**:39.
61 Khosravi-Maharlooei M, Obradovic A, Misra A, Motwani K, Holzl M, Seay HR, *et al.* Cross-reactive public TCR sequences undergo positive selection in the human thymic repertoire. *J Clin Invest.* 2019; **129**:2446–62.
62 DeWitt WS, Smith A, Schoch G, Hansen JA, Matsen FA, Bradley P. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *eLife* 2018; **7**:e38358.
63 Redmond D, Poran A, Elemento O. Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. *Genome Med.* 2016; **8**:80.
64 Sun X, Saito M, Sato Y, Chikata T, Naruto T, Ozawa T, *et al.* Unbiased analysis of TCRα/β chains at the single-cell level in human CD8+ T-cell subsets. *PLoS One* 2012; **7**:e40386.
65 Turchaninova MA, Britanova OV, Bolotin DA, Shugay M, Putintseva EV, Staroverov DB, *et al.* Pairing of T-cell receptor chains via emulsion PCR: new technology. *Eur J Immunol.* 2013; **43**:2507–15.
66 Howie B, Sherwood AM, Berkebile AD, Berka J, Emerson RO, Williamson DW, *et al.* High-throughput pairing of T cell receptor α and β sequences. *Sci Transl Med.* 2015; **7**:301ra131.
67 Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 2017; **547**:89–93.
68 Cukalac T, Kan W-T, Dash P, Guan J, Quinn KM, Gras S, *et al.* Paired TCRαβ analysis of virus-specific CD8 $^{+}$ T cells exposes diversity in a previously defined "narrow" repertoire. *Immunol Cell Biol.* 2015; **93**:804–14.
69 Petrie HT, Livak F, Schatz DG, Strasser A, Crispe IN, Shortman K. Multiple rearrangements in T cell receptor alpha chain genes maximize the production of useful thymocytes. *J Exp Med.* 1993; **178**:615–22.
70 Grigaityte K, Carter JA, Goldfless SJ, Jeffery EW, Hause RJ, Jiang Y, *et al.* Single-cell sequencing reveals αβ chain pairing shapes the T cell repertoire. *bioRxiv* 2017;213462. https://doi.org/10.1101/213462
71 Dupic T, Marcou Q, Walczak AM, Mora T. Genesis of the αβ T-cell receptor. *PLoS Comput Biol.* 2019; **15**:e1006874.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Empirical cumulative distribution of nucleotide read abundances, with 100 bootstrapped replicates for individual mouse samples. Each individual was resampled independently in the bootstrap replicates based on individual read abundances. Cumulative distributions of abundances in individual mice showed similar trends to the pooled data. As sharing level increases in the productive repertoire, the distribution shifts towards higher abundances, but the opposite trend occurs in the nonproductive repertoire.

Figure S2. (A) ECDF for amino acid TCRβ sequences with bootstrap resampling. (B) Proportion of unique amino acid clonotypes that are encoded by x unique nucleotide sequences, divided by sharing level and functional status. The vast majority of amino acid sequences are encoded by only one nucleotide sequence in the private repertoires, but as sharing increases, the mode of the number of degenerate nucleotide sequences per amino acid sequence increases.

Figure S3. ECDFs of productive and nonproductive nucleotide sequences filtered for both high and low-abundance sequences and with productive sequences downsampled to the size of the nonproductive sequences show abundance shifts consistent with observations of the full data. In (A), ECDFs were drawn from 100 bootstrap samples excluding any sequence with an abundance smaller than 10 copies, therefore removing potentially error-prone reads. In (B), ECDFs were drawn from 100 bootstrap samples excluding any sequence with an abundance larger than the 95th percentile in each mouse, removing potential memory cell contaminants. Both tests show that the shifts of abundances are not driven by reads originating from potential memory T cell contamination nor from erroneously sequenced unique reads.

Figure S4. Inferred generation probability does not fully explain nucleotide sequence abundance nor sharing proportion. (A) shows the correlation between TCRβ abundance and inferred generation probability, split by functional status and mouse. Plotted with log-transformed axis and simple linear regression on transformed data. (B) shows the proportion of sequences that are shared, binned by generation probability such that each bin has approximately the same number of unique sequences (approximately 2600 per nonproductive bin and 6200 per productive bin). 95% proportion confidence intervals were estimated by Wilson's method. Sharing proportions increase with generation probability.

Figure S5. Heat map of sum of squared deviation of gene usage proportions of unique sequences between subsets of the repertoire for (A) V gene segments and (B) J gene segments. (C) shows the distribution of V gene segment usage for unique sequences in the productive and nonproductive repertoire, split by sharing level.

Figure S6. Inferred public fraction of receptors for repertoire sizes ranging from 10^7 to 10^9, calculated from generation probability distributions of (A) four subsets the sampled repertoire, and (B) productive and nonproductive subsets as a whole. We used the expression derived in Elhanati et al., 2018 for the expected public fraction, computing the density of sequences in each bin of generation probability from our four (or two in panel B) empirical distributions of generation probability and midpoint approximation of the integral. We downsampled the generation probability distributions for the productive sequences to the size of the nonproductive sequences and calculated the public fraction for 100 bootstrap replications.

Table S1. Number of unique (and total) TCRβ sequences per mouse. Nonproductive amino acid sequences included only in-frame (stop codon) sequences.